Spectrometry Classification Algorithms (*MSCA*)

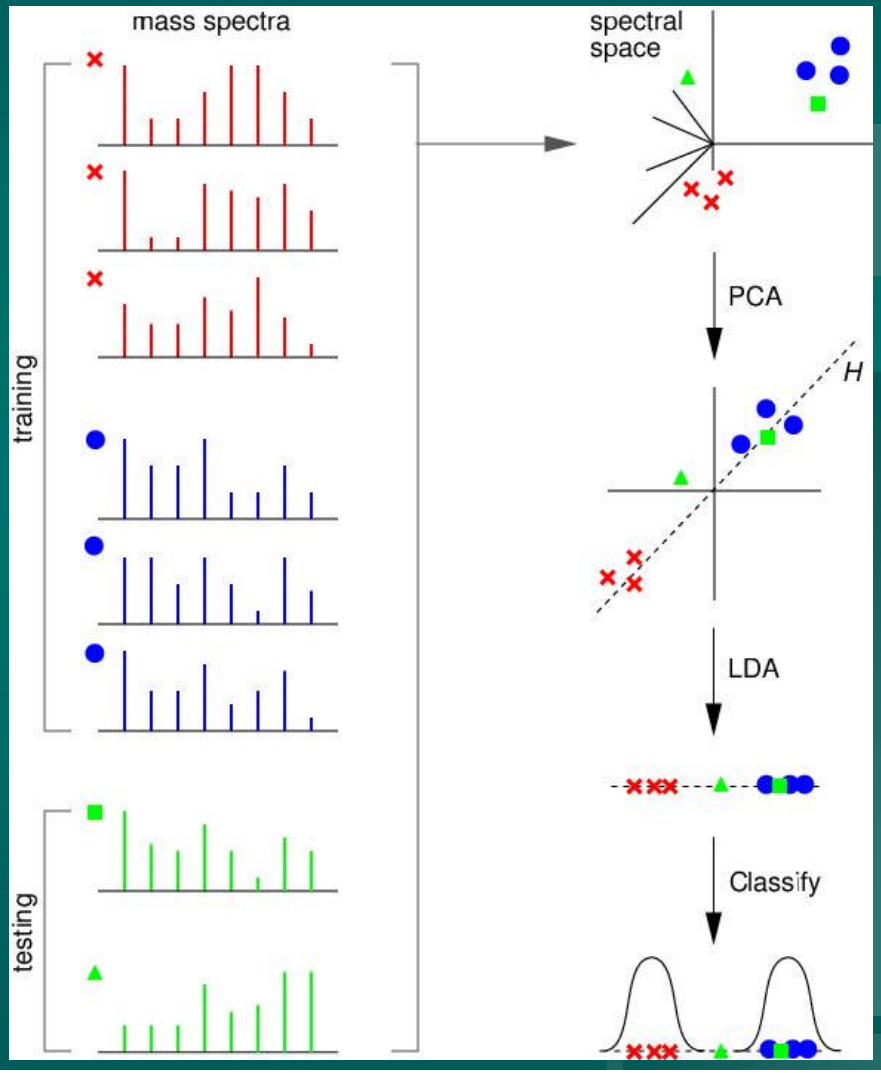
Disease v.s. Healthy

(1.) Do differences exist between the two states?

(2.) What molecules fit within the *m/z* and what are the identities of the contributing molecules?

Q5

Disease classification by mass spectrometry pattern recognition



Proteome Analysis

Expression analysis: proteins (mass spectrometry)

Algorithm uses PCA followed by LDA

probabilistic classification of healthy vs. disease whole serum samples using mass spectrometry

Q5 is a closed-form exact solution for classification of complete Mass Spectrometry

Q5 is cast as a MSCA that is closed-form (can be solved using singular value decomposition) (non-iterative and deterministic)

It is combinatorially precise:

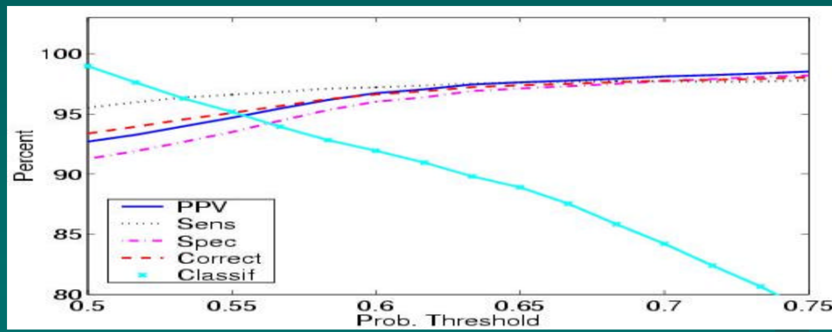
(1.) Run time (or *complexity*) can be computed

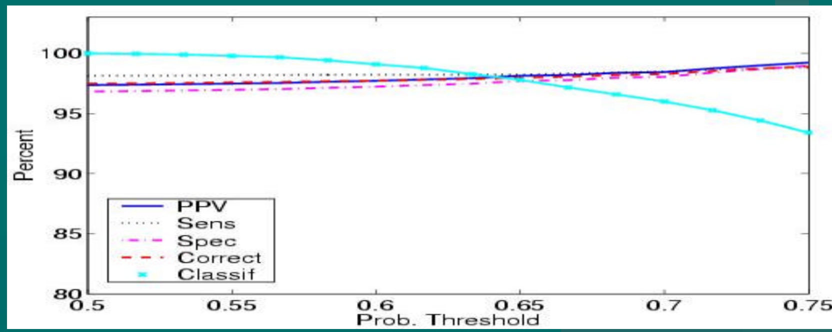
$T_{tr}=O(n^3+n^2r)$ and $T_{test}=O(mrn)$

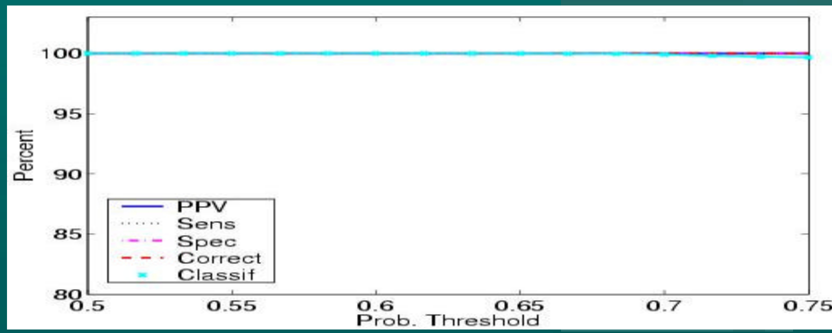
(2.) Always provides and optimal solution (or *correctness*) to an objective error function

Q5

Disease classification by mass spectrometry pattern recognition







Both SELDI protein chip type and sample preparation method contribute significantly to classification accuracy.

Achieve sensitivity, specificity, and positive predictive values above 97% on three ovarian cancer datasets and one prostate cancer dataset

3-class classification results

PC-IMAC-Cu dataset (0.63 prob-thres)

50%

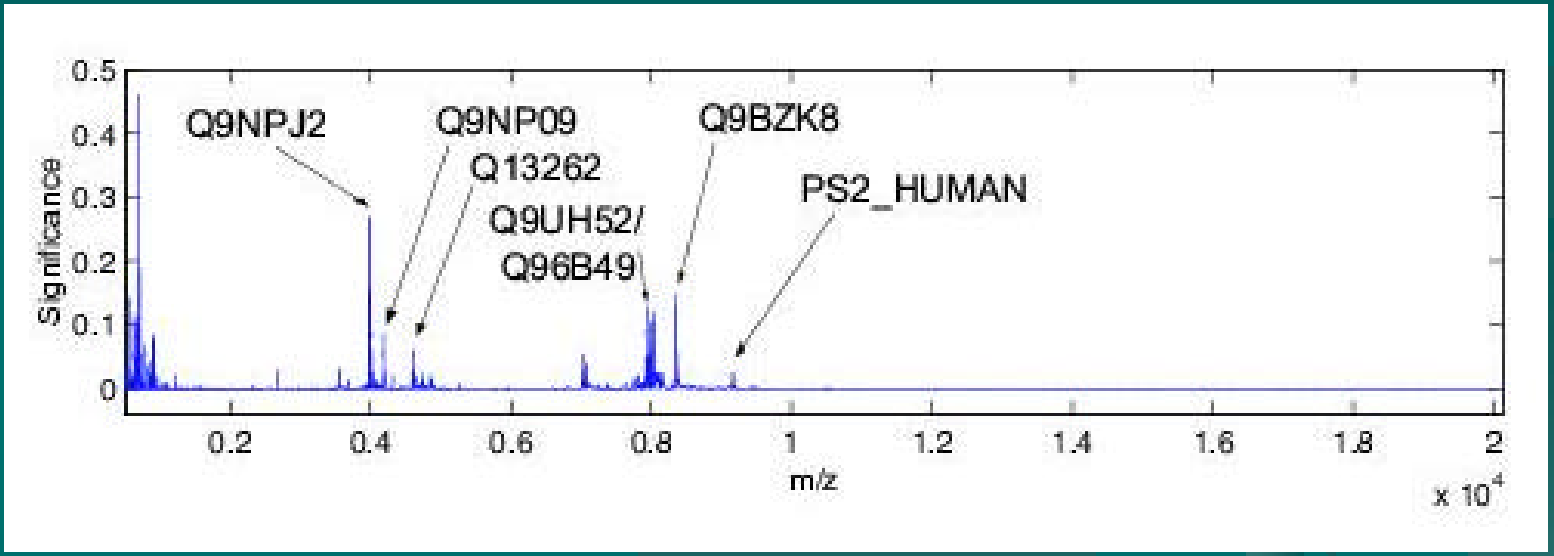
Spectra Type	NH	BPH	PC
NH	99.9(0.5)	0.0(0.0)	0.1(0.5)
PBH	0.1(0.4)	91.0(6.2)	8.9(6.2)
PC	0.4(0.8	4.0(2.7)	95.2(2.8)

95%

Spectra Type	NH	BPH	PC
NH	100.0(0.0)	0.0(0.0)	0.0(0.0)
PBH	0.0(0.0)	95.2(13.2)	4.2(12.5)
PC	0.1(1.4)	3.6(7.2)	96.3(7.8)

Q5

Disease classification by mass spectrometry pattern recognition



These SWISSPROT and TrEMBL proteins are consistent with *m/z* peaks of the discriminant having significance for classification of Ovarian Cancer serum samples. Due to mass-aliasing, the database lookup does not prove these proteins present in the serum samples, but these proteins serve as candidates in the search for novel biomarkers.

Proposed applications to develop further with full caBIG API interoperability

Open access application ?	Q5
-source code available	+
-use cases are reported (in press)	+

-ported and wrapped with open source technologies for public access

-support metadata development with database curating activities via caBIG APIs

References

Q5 Lilien, R.H., Farid, H., Donald, B.R. 2003. Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. J. Comput. Biol. 10

http://www.cs.dartmouth.edu/~donaldlab/Software/